Graduate Education

Interpreting the literature in obstetrics and gynecology: I. Key concepts in epidemiology and biostatistics

Herbert B. Peterson, MD, and
David G. Kleinbaum, PhD

The proper interpretation of research findings in obstetrics and gynecology increasingly requires some understanding of epidemiology and biostatistics. The disciplines of epidemiology and biostatistics are inextricably related; the goal of epidemiology is accurate measurement of the relationship between an exposure and a disease, and statistical methods are required for achieving that objective. Most epidemiologic studies in the obstetrics and gynecology literature can be classified as 1) cross-sectional, 2) case-control, or 3) cohort (follow-up) studies. The 2 × 2 table represents the basic analytic format for all three types of epidemiologic studies. Information from this table can be used to estimate both the magnitude of the exposure-disease relationship and the relative likelihood that chance explains study findings. Accurate measurement of the relationship between an exposure and a disease can be impeded by two major sources of error: bias and chance. In broad terms, biases can be classified as those related to 1) selection, 2) information, and 3) the presence of extraneous variables. Because biases in epidemiologic studies distort measurements, they must be identified, characterized, and, if possible, avoided. When biases cannot be avoided, knowledge of their likely impact on study findings must be assessed. The role of chance is evaluated by statistical testing of the null hypothesis, ie, the hypothesis that two factors are not associated. Statistical significance is only one consideration in the evaluation of study findings; to determine whether an observed association is likely to be important clinically, the critical reader needs to go beyond chance (*P* values) to consider other important criteria, including strength of the association, consistency of the study findings with known information, and biologic plausibility of the observed association. *(Obstet Gynecol 78:710, 1991)*

Scientific publications and presentations provide information regarding new developments in obstetrics and gynecology, but proper interpretation of this information increasingly requires some understanding of epidemiology and biostatistics. In this overview, we discuss principles that should help the clinician to interpret the literature in obstetrics and gynecology.

Epidemiology

Many definitions of epidemiology have been offered; a useful one is "the study of disease and health in human populations."[1] The objective of an epidemiologic study is to measure the relationship between an exposure of interest (eg, ingestion of a drug) and an outcome of interest (eg, occurrence of a disease).

Epidemiologists agree that causation cannot be proven from a single study; it can only be inferred from the aggregate results of several studies. Nevertheless, certain types of research designs provide a stronger basis for causal inference than others. For example, experiments are considered "stronger" than nonexperimental analytic studies because the exposure of interest can be manipulated; the exposure's impact on the outcome can thus be directly estimated. Randomized clinical trials, in which the participants undergo various treatment regimens for a particular disease, are the major examples of experimental epidemiologic studies in our literature. Experimental studies in humans are generally difficult to conduct, however, and are sometimes ethically unacceptable. Therefore, most of the epidemiologic studies in our literature are nonexperimental, ie, observational. In observational studies, the exposures of interest are observed rather than manipulated or randomized.

From the Division of Reproductive Health, Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control, Atlanta, Georgia; and the Departments of Biostatistics and Epidemiology, School of Public Health, University of North Carolina, Chapel Hill, North Carolina.

		Exposed	
		Yes	No
Diseased	Yes	a	b
	No	c	d

Figure 1. A 2 × 2 table that represents the basic analytic format for observational studies. a, b, c, d = the numbers of people in each of four possible combinations of exposure and disease status.

Observational Study Designs

Most of the observational epidemiologic studies in the obstetrics and gynecology literature can be classified as either 1) cross-sectional, 2) case-control, or 3) cohort (follow-up) studies. The distinguishing features of these studies are generally easily recognized and often quite important. One useful way to appreciate the similarities and differences among the three types is to consider the simple case of one exposure variable and one outcome variable. The relationship between an exposure and a disease can be determined from information contained in a 2 × 2 table (Figure 1). The four letters (a, b, c, and d) represent respective counts of study subjects falling in one of the four possible exposure-disease combinations. For example, the "a" combination includes those subjects who were both exposed and diseased, whereas the "d" combination includes those subjects who were neither exposed nor diseased.

The 2 × 2 table represents the basic analytic format for all three types of analytic epidemiologic studies. Although a detailed description of these three types of studies is beyond the scope of this presentation, we can discuss some simple but important features of each.

Cross-sectional studies evaluate populations of individuals, some of whom may have the disease (outcome) of interest and some of whom do not. This type of study can be thought of as a snapshot of a group of people characterizing them by whether they do or do not have the disease of interest at one particular moment and by whether they are or are not exposed to the factor of interest at that moment. Individuals who have the disease at that moment are considered "prevalent" cases. The difference between prevalent cases and "incident" cases is important in this context. Incident cases are new cases of disease occurring over a specific period of time. Use of incident cases, as compared with prevalent cases, generally permits stronger conclusions regarding the likelihood that an exposure is causally related to a disease. Thus, cross-sectional studies, which by design use prevalent cases, may not permit strong arguments for causation. For example, a cross-sectional study to evaluate the relationship between condom use and human immunodeficiency virus (HIV) transmission would characterize persons by whether they had HIV infection at a particular time and by whether they used condoms at that time. Unfortunately, because the cross-sectional design uses prevalent cases only, the temporal association between condom use and acquisition of HIV infection cannot be determined in this study. Thus, individuals who had consistently and correctly used condoms may have done so after having become infected with HIV, and individuals who had correctly used condoms at some time in the past may have stopped using them for a brief period, become infected, and subsequently resumed correct use. Because a cross-sectional study cannot assess the temporal relationship between condom use and HIV infection, it provides limited and potentially misleading information regarding the etiologic relationship between condom use and HIV transmission.

Case-control and cohort studies differ from cross-sectional studies in that they both can include the experience of incident cases. Case-control studies can include either prevalent or incident cases. Cohort studies, by design, allow only incident cases. The major difference between case-control studies and cohort studies is that case-control studies classify study subjects on the basis of whether they are diseased and then determine whether they were previously exposed to the factor of interest. Cohort studies, on the other hand, classify study subjects on the basis of exposure status and then follow them to determine whether they develop disease. In the example of condom use and HIV transmission, a case-control study would identify participants as either prevalent or incident cases of HIV infection and then determine whether the participants had or had not used condoms recently or in the past. By contrast, a cohort study would identify participants by whether they used condoms and then follow the subjects over time to determine whether they develop HIV infection. Certainty about an individual's HIV infection status would be contingent upon identifying individuals known to be uninfected (seronegative) at one point in time and known to have become infected (seroconverted) at a subsequent time.

Case-control studies are retrospective: Study subjects are first identified as being diseased or not diseased; the investigator then determines whether the study subject was exposed sometime in the past. Cohort studies, by contrast, typically involve identifying a population of disease-free individuals who are either exposed or not exposed and then following them prospectively to see whether they develop disease. Thus, cohort studies are often called prospective studies. Because retrospective cohort studies are also possible, however, many epidemiologists use the term "follow-up" or "longitudinal" when referring to those cohort studies that follow individuals over time to determine whether they develop disease. For the rest of this overview, such cohort studies will be referred to as follow-up studies.

Case-control studies have several advantages over follow-up studies. First, they can be completed in less time than follow-up studies, particularly when the disease of interest has a long induction period. Follow-up studies require that subjects be followed for at least as long after exposure as during the known or suspected induction period, which for some diseases (such as cancer and cardiovascular disease) may be 10–20 years or more after exposure. Such long-term studies are difficult to conduct and generally quite expensive. Further, a large number of study subjects may be lost to follow-up during that period, making the interpretation of study results difficult or impossible. Thus, case-control studies are typically more efficient and less costly than follow-up studies. This is particularly true when the disease being studied is uncommon. For example, hepatocellular adenoma occurs in fewer than four of every 100,000 oral contraceptive (OC) users. A follow-up study to assess the relationship between OC use and hepatocellular adenoma would therefore have to follow at least 100,000 OC users for years before even a few cases of hepatocellular adenoma would be expected. If such a study followed substantially fewer than 100,000 OC users and identified no cases of hepatocellular adenoma, the negative finding would be uninterpretable. It would be much more efficient to a identify a group of women with this rare condition and a control group of disease-free women and then determine whether the members of either group had previously used OCs.

Despite these advantages, case-control studies are sometimes considered methodologically inferior to follow-up studies. In follow-up studies, data are collected forward, from exposure toward effect. By contrast, in case-control studies, data are collected backward, from effect toward exposure. Nevertheless, a "perfectly" done case-control study should provide as accurate a characterization of what is being measured as a "perfectly" done follow-up study. As a practical matter, however, no epidemiologic study is ever perfect—all studies have at least some methodologic limitations. These limitations are inherently greater in case-control studies. In particular, accurate documentation of exposure history can be very difficult. Often, limitations of case-control studies are related to difficulties in selecting proper controls. Persons in the control group of a case-control study should meet strict criteria, including the following: being at risk of developing the disease under study, being as comparable to cases as possible (except for having the disease), and not having other conditions related to the likelihood of having the exposure of interest. Despite biases often inherent in the measurement of past exposure and the selection of study controls, a well-designed and well-conducted case-control study may have substantially fewer methodologic limitations than a poorly designed and poorly conducted follow-up study.

Precision Versus Validity

Accurate measurement of the relationship between an exposure and a disease can be impeded by two major sources of error: random error and systematic error. Precision corresponds to random error and validity to systematic error. Specifically, precision refers to the extent to which random error or chance affects the results of one's study. The more precise a study, the less likely its findings are attributable to chance. An epidemiologic investigation is conducted on samples of people, the inclusion of whom is determined by chance; therefore, the results of analysis in two or more samples may differ, purely by chance. In general, the larger the study population (study size), the greater the precision.

In contrast to precision, validity concerns whether there is a systematic error, in either the research design or the analysis, that leads to a wrong conclusion. In particular, the data resulting from a poor study design may suggest a strong association between the study exposure and the disease when, in fact, there is no association at all. Conversely, the data may indicate no association when, in reality, a strong association exists. A distortion that may result when estimating the association of interest is usually called a bias. Bias has been defined as "any effect of any stage of investigation or inference tending to produce results that depart systematically from the true values."[2] Bias can result from the way subjects are selected into the study, from incorrect information gathered on study subjects, and from failure to adjust for variables (other than the exposure) that may influence the likelihood of becoming diseased.

Validity can be clarified (and contrasted with precision) by using the marksman's target as a metaphor. Precision can be thought of as the proximity of a shot, which represents the observed study sample, to the bull's-eye of the target being fired at, ie, how close to hitting the target the investigator comes. Validity, by contrast, deals with whether the shot is being fired at the right target. When a study is invalid, it is shooting at the wrong target and consequently is not measuring what it is supposed to measure. In further distinguishing between validity and precision, note that it is possible to get a very precise estimate that gives a biased or wrong conclusion (the shots are closely grouped but missing the correct target). Generally, when doing epidemiology studies, it is important not to sacrifice validity for the sake of precision; the goal is to get the right (unbiased) answer, rather than get a precise answer that is nevertheless misleading.

Biases

Because biases in epidemiologic studies distort measurements and may lead to a wrong answer, they must be identified, characterized, and, if possible, avoided. When biases cannot be avoided, knowledge of their likely impact on the distortion of results will greatly aid in the interpretation of study findings. If biases are not so characterized, the study results may be uninterpretable. In broad terms, the numerous potential study biases can be classified as those related to 1) selection, 2) information, and 3) the presence of extraneous variables.

Selection bias is a measurement error attributable to the procedure for selecting study participants. It results in measures of effect different from those that would be obtained if the entire target population were studied. There are many types of selection bias. One type is due to self-selection: Study participants who volunteer for an investigation may be more or less likely to have the exposure of interest; for example, women who took a suspected teratogen and whose infants had a birth defect may be more likely than other women to volunteer for a study of birth defects. A second type of selection bias concerns follow-up of study participants. If subjects exposed to the factor of interest were more likely to be followed than were nonexposed persons, such unequal follow-up might bias the study toward a positive association between exposure and disease, particularly if those followed were more likely to get the disease than those not followed. Selection bias can also occur in the selection of study controls. Comparison groups should be as similar as possible to the case or exposed group. And as noted earlier, in a case-control study, the control group should be at risk for having the disease under study and should not have conditions related to the exposure of interest.

Information bias, or misclassification bias, results from systematic errors in the collection of information about either the exposure or the disease being evaluated in the study. If the error for either the exposure or the disease is independent of the other factor, the error is termed nondifferential misclassification. If the error for the exposure is not independent of the disease or vice versa, the error is termed differential misclassification. One example of differential misclassification is recall bias, which can result from selective recall of study subjects. Persons with the disease of interest may be more likely than those without the disease to recall exposures that they consider related to their disease. Women whose children had birth defects may be more likely than are women whose children did not have birth defects to recall a variety of exposures, including ingestion of a teratogen. Nondifferential error results in bias toward an underestimation of any real effect. In contrast, differential misclassification can lead to a bias that either overestimates or underestimates the true effect.

Measurement of the exposure-disease relationship can be further complicated by extraneous factors called covariates. To understand how covariates may introduce measurement errors, it is important to understand the terms "interaction" and "confounding."

Interaction and Confounding

Interaction occurs when the relationship between the exposure and the disease being measured varies according to the level (ie, value) of one or more covariates.[3] When such variation occurs, the covariate is considered an effect modifier. As an example, the relationship between OC use and cardiovascular disease varies depending on whether a woman smokes. Oral contraceptive users who smoke are at greater relative risk of myocardial infarction than those who do not smoke. Because the relative risk differs for smokers and non-smokers, interaction of the exposure (OC use) with smoking is said to be present. When significant interaction occurs, as determined by statistical testing, relative risk estimates (to be discussed later) should be presented separately for those with and those without the modifying covariate (effect modifier), in this case, smoking. The importance of separate relative risk estimates is obvious in the stated example; the separate risks identified have implications for both clinical decision-making and counseling.

Confounding, in simple terms, is the mixing of effects.[4] It results in an inaccurate measure of the effect of an exposure if the extraneous factor, or confounder, is not taken into account in the analysis. To be a confounder, the extraneous factor must be associated with both the likelihood of being exposed and the likelihood of developing the disease. When confounding is present, an estimate of the relationship between the exposure and the disease that does not account for a suspected confounder will be meaningfully different from an estimate that adjusts for the suspected confounder using special analytic techniques. For example, if we found a strong relationship between OC use and myocardial infarction when smoking was ignored in the analysis and a weak or absent relationship when smoking was considered, we would conclude that smoking is a confounder. Unlike interaction, which, as noted, is assessed by statistical testing, confounding is evaluated without statistical testing. Adjusting or controlling for confounding is discussed in more detail in Part II of this report.

Biostatistics

Our distinction between epidemiology and biostatistics is arbitrary; actually the two disciplines are inextricably related. Indeed, some important principles of biostatistics have already been introduced here. Statistical methods are required for reaching epidemiology's goal of accurate measurement of an exposure-disease relationship.

Testing Hypotheses

Epidemiology tests hypotheses. By convention, statistical methods test the null hypothesis, ie, the hypothesis that two factors are not associated. If study findings indicate that the hypothesis of no association can be rejected, then the alternative hypothesis that some association exists is accepted. The association may be large or small, biologically meaningful or not. The decision to reject or not reject the null hypothesis is usually based on the *P* value. The *P* value indicates the relative likelihood that the observed exposure-disease relationship is due to chance. Typically, $P < .05$ is used to determine rejection of the null hypothesis. Such rejection means that the observed association between two factors is unlikely (less than 5% likelihood) to be due to chance. In other words, there is less than a 5% chance that the decision to reject the null hypothesis is in error. When the null hypothesis is rejected using $P < .05$, we say that the observed association is statistically significant at the 5% level.

Errors known as type I and type II can occur when the null hypothesis is tested. A type I error occurs when the null hypothesis is true but is incorrectly rejected, ie, concluding incorrectly that two factors are associated. The likelihood of a type I error is fixed when the significance level is chosen. For example, if the significance level of .05 is chosen, then 5% of the time the null hypothesis will be rejected when it should not be. The significance level and the *P* value are not the same. The significance level is chosen by the investigator, ideally without looking at the data, and is the probability of making a type I error that the investigator is willing to allow. The *P* value, by contrast, is a so-called posterior probability value based on the data; it represents the likelihood of observed differences under the null hypothesis. This likelihood (ie, the *P* value), because it is based solely on the observed data, may be either higher or lower than what the investigator is willing to allow based on a predetermined significance level.

A type II error occurs if the null hypothesis is not rejected when it should be, ie, concluding incorrectly that two factors are not associated. To understand this type of error, one needs to understand the concept of study power. In simple terms, a study's power is its ability to significantly detect an association if it really exists. This ability is contingent on study size: In general, the larger the study population, the greater the study power. A type II error can occur when a study is too small to detect an association that really exists. For example, suppose maternal exposure to a drug causes birth defects in 1.0% of infants. In one study, the rate of anomalies among the infants of 50 women exposed to the drug is compared with that among infants of 50 women not exposed. No anomalies were reported (and given the known 1% rate at which the drug causes birth defects, none should have been expected). The conclusion that maternal exposure to the drug is not associated with birth defects might be a type II error because the sample size of 50 in each group may have been too small to detect a real effect. Thus, when results indicate no association between factors being evaluated, we should determine whether the study is large enough to have had the potential to detect the association.

The *P* value expresses the relative likelihood that chance explains study findings. Chance, however, is only one factor to consider when determining whether an observation is likely the result of cause and effect. To try to determine causation, epidemiologists also assess the strength of an observed association. Both case-control and follow-up studies measure the magnitude of association between the exposure and the disease of interest. In follow-up studies, the strength of the association can be estimated by using the relative risk. In case-control studies, the odds ratio is typically used. With rare diseases, the odds ratio from a case-control study with proper controls closely approximates the relative risk from a follow-up study. The numerical value obtained for either a relative risk or an odds ratio can be interpreted in the same way. In essence, each measure compares the risk for exposed persons with the risk for unexposed persons. For example, a relative risk or odds ratio of 1 means that the risk for exposed persons is the same as that for unexposed persons, ie, there is no association between exposure and disease. When the relative risk or odds ratio is greater than 1, the risk (or ratio of the relative odds) is greater for exposed persons than for unexposed persons. In this situation, we say that the direction of the association is positive. For example, if the estimate is 10, the risk for exposed persons is ten times greater than for unexposed persons. If the relative risk or odds ratio is less than 1, the risk is lower for exposed persons than for unexposed persons. In this situation, we say that the direction of the association is negative. For example, if the estimate is 0.1, the risk for exposed persons is one-tenth that for unexposed persons.

Calculation of both the relative risk and the odds ratio requires the information from the classic 2 × 2 table. As already noted, the table (Figure 1) is based on information, obtained in an observational study, regarding those exposed and those unexposed to the factor of interest and those known to be diseased and those not diseased. With this information, the estimated relative risk (Figure 2) and the estimated odds ratio (Figure 3) can be calculated.

		Exposed	
		Yes	No
Diseased	Yes	a	b
	No	c	d
		N_1	N_0

$$\text{Relative risk} = \frac{\text{risk for the exposed}}{\text{risk for the unexposed}} = \frac{a/N_1}{b/N_0} = \frac{a \times N_0}{b \times N_1}$$

Figure 2. Calculation of the relative risk using the 2 × 2 table. a, b, c, d = the numbers of people in each of four possible combinations of exposure and disease; N_1 = total number exposed; N_0 = total number unexposed.

		Exposed	
		Yes	No
Diseased	Yes	a	b
	No	c	d

$$\text{Odds ratio} = \frac{\text{exposure odds among cases (diseased)}}{\text{exposure odds among controls (not diseased)}}$$

$$= \frac{\frac{a}{a+c}}{\frac{b}{b+d}} = \frac{\frac{a}{c}}{\frac{b}{d}}$$

$$= \frac{ad}{bc}$$

Figure 3. Calculation of the odds ratio using the 2 × 2 table. Abbreviations as in Figure 1.

Confidence Intervals

The observed odds ratio and relative risk are reported as "point estimates." A confidence interval for this estimate indicates the variability of the point estimate. The wider the confidence interval, the larger is the variability of the point estimate and the less likely that the point estimate is accurate. The 95% confidence interval is often used. Technically, this means that if the study were to be repeated over and over again and a 95% confidence interval were calculated each time, 95% of these confidence intervals would be expected to contain the true exposure-disease parameter being estimated. Note that it would be incorrect to characterize the confidence interval as giving a range of values for the "true" exposure-disease value; the true value is a single number that does not vary.

The confidence interval for the point estimate can also be used to determine statistical significance for a two-tailed significance test, in which the investigator allows for the possibility that the exposure either increases or decreases the risk of the disease. Generally, if α is a (preset) significance level, then the null hypothesis is rejected if the 100 $(1 - \alpha)$% confidence interval does not overlap the null value being tested. For example, if the measurement is a relative risk, the null value is relative risk = 1. Then, if a 95% confidence interval for the relative risk does not overlap 1.0, the null hypothesis of no exposure-disease association is rejected at the .05 significance level. The confidence interval can therefore be used instead of the P value to test the likelihood that study results are due to chance.

Table 1. Suggested Readings

Greenberg RS, Kleinbaum DG. Mathematical modeling strategies for the analysis of epidemiologic research. Ann Rev Public Health 1985;6:223–45.

Hill AB. The environment and disease: Association or causation? Proc R Soc Med 1965;58:295–300.

Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research: Principles and quantitative methods. New York: Van Nostrand Reinhold, 1982.

Kleinbaum DG, Kupper LL, Muller KE. Applied regression analysis and other multivariable methods. Boston: PWS-Kent, 1988.

Last JM. A dictionary of epidemiology. New York: Oxford University Press, 1983.

Rothman KJ. Modern epidemiology. Boston: Little, Brown, 1986.

Schlesselman JJ. Case-control studies: Design, conduct, analysis. New York: Oxford University Press, 1982.

For example, if the 95% confidence interval is 2.1–9.5, this does not overlap 1.0, so that the null hypothesis is rejected at the 5% significance level; this is the same conclusion we would reach if $P < .05$. However, if the 95% confidence interval is 0.5–9.5, this overlaps 1.0, so that the null hypothesis is not rejected, which is the same conclusion we would reach if $P > .05$.

Mathematical Modeling

The analysis of epidemiologic data typically requires the use of complex statistical procedures involving mathematical modeling. The most commonly used mathematical model in the literature on obstetrics and gynecology is logistic regression. In Part II of this report, we describe how logistic regression can control for multiple confounders.

Interpreting the Literature

How can this brief discussion help the practicing clinician better interpret the literature? Each point discussed here relates to questions the critical reader must ask. Here are a few examples:

1) What exposure-disease relationship is being studied or what hypothesis is being tested? The most important question the critical reviewer should initially ask is, "What is being measured?"

2) Is the study population appropriate for testing the hypothesis?

3) Is the study methodology appropriate for testing the hypothesis? What type of study was conducted, and what are its major methodologic limitations?

4) What factors other than the exposure and disease under study need to be considered? What study biases may apply? Have they been adequately identified and controlled? If not, how should the results be interpreted?

5) If no association was found, was the study power adequate to detect an existing important association?

By asking these questions, the reader can determine whether observed associations are likely due to chance, selection or information bias, confounding (some epidemiologists consider that confounding is a bias as well), or cause and effect. When associations are based on studies that have serious methodologic limitations, results are difficult or impossible to interpret.

For clinicians, the primary question is whether observed associations are important clinically. *P* values tell us whether an observed association is likely due to chance, but they often tell us nothing about clinical relevance. Clinical relevance is probably contingent upon whether an observed association is causal and, if causal, upon the direction and magnitude of the observed relationship expressed by odds ratio and relative risk estimates. Although epidemiologic studies cannot, strictly speaking, prove causation, they can be used to infer causation when properly conducted and interpreted. To determine the likelihood of causation, the critical reader must go beyond chance (*P* values) to consider other criteria, including the following[5]:

1) Strength of the observed association. In general, the stronger the association, the more likely it is to be real.

2) Consistency of the study findings with those of other reports and all known information about the exposure and the outcome investigated.

3) Temporality of the observed association. Does the cause precede the effect?

4) Biologic plausibility of the observed association. Does the relationship make sense?

5) Biologic gradient in the observed association. Is there a dose-response relationship?

6) Coherence with what is known regarding the natural history and biology of the outcome under study.

7) Experimental evidence to support or refute the observed association.

8) Analogy. Is the observed association supported by similar associations?

Although this list of criteria is incomplete and cannot always be used to establish probable cause, these and other factors must be considered to determine whether statistically significant findings are clinically significant as well.

A basic understanding of epidemiology and biostatistics is essential for proper interpretation of the literature in obstetrics and gynecology. The principles discussed here, supplemented by appropriate texts (Table 1), should equip the clinician with most of the tools necessary for the job. A dictionary of epidemiologic terms is available (Table 1) and may serve as a useful reference. Like other tasks, reviewing the literature gets easier with practice. Despite appearances, one does not need to be an epidemiologist to properly interpret most analytic epidemiologic studies; however, consultation with an epidemiologist may be useful, particularly to address concerns beyond the scope of this discussion.

References

1. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research: Principles and quantitative methods. New York: Van Nostrand Reinhold, 1982.
2. Last JM. A dictionary of epidemiology. New York: Oxford University Press, 1983.
3. Greenberg RS, Kleinbaum DG. Mathematical modeling strategies for the analysis of epidemiologic research. Ann Rev Public Health 1985;6:223–45.
4. Rothman KJ. Modern epidemiology. Boston: Little, Brown, 1986.
5. Hill AB. The environment and disease: Association or causation? Proc R Soc Med 1965;58:295–300.

Address reprint requests to:
Herbert B. Peterson, MD
Women's Health and Fertility Branch
Centers for Disease Control, Mailstop K–34
1600 Clifton Road
Atlanta, GA 30333

Received November 26, 1990.
Received in revised form April 23, 1991.
Accepted April 23, 1991.

Interpreting the literature in obstetrics and gynecology: II. Logistic regression and related issues

Herbert B. Peterson, MD, and
David G. Kleinbaum, PhD

The goal of epidemiology is accurate measure of the relationship between an exposure and a disease of interest. The control of covariates of the exposure-disease relationship is required to obtain a valid measure. Two types of covariates, confounders and effect modifiers, must be considered. Investigators can design studies to measure and control for the impact of both types of covariates. Design strategies for dealing with covariates include randomization, restriction, and matching. If the impact of a covariate is not eliminated by study design, it must be controlled for during study analysis by use of either stratification or mathematical modeling. Stratified analysis permits an assessment of the exposure-disease relationship for each category of relevant covariates. Although stratification is the best initial approach for controlling covariates, it is often impractical, particularly if more than one or two covariates must be controlled. Multivariate mathematical models are required if multiple covariates are to be controlled. Logistic regression is the mathematical modeling procedure most often used to analyze studies in obstetrics and gynecology. Although there are no uniform rules for building a proper model for regression analysis, useful general strategies are available. It must be emphasized that, though the use of mathematical modeling can control for multiple covariates and thereby improve the chance of obtaining an accurate measure of the exposure-disease relationship, it cannot "fix" data that result from a poorly designed or improperly conducted study. *(Obstet Gynecol 78:717, 1991)*

In Part I of this report, we highlighted important principles of epidemiology and biostatistics relevant to interpreting the literature in obstetrics and gynecology. These principles included the control of covariates of the exposure-disease relationship. Two types of covariates were distinguished: confounders and effect modifiers; both are extraneous to the exposure-disease relationship and must be considered to obtain a valid measure of that relationship. In Part II, we discuss the various strategies investigators use to control for the effects of covariates in both the design and the analysis of a study. In particular, we discuss how several covariates can be controlled simultaneously by using a mathematical modeling procedure called logistic re-

From the Division of Reproductive Health, Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control, Atlanta, Georgia; and the Departments of Biostatistics and Epidemiology, School of Public Health, University of North Carolina, Chapel Hill, North Carolina.

0029-7844/91/$3.50

gression, which is used increasingly to analyze studies in obstetrics and gynecology. Although a complete discussion of logistic regression usually requires a semester or more in graduate school, we highlight the subject here with a minimum of mathematics and jargon.

Dealing With Covariates Through Study Design

Before designing a study, an investigator should identify potential covariates associated with the exposure-disease relationship under investigation by reviewing the relevant literature. Once such covariates are identified, the investigator can design the study to measure and control for their impact.

Design strategies for dealing with covariates include randomization, restriction, and matching. Except for randomization, which can only be used for experimental studies, these techniques can be applied to both experimental and nonexperimental investigations.

To visualize randomization, suppose that the relationship between oral contraceptive (OC) use and myocardial infarction is to be measured and that smoking is considered to be a covariate (see Part I). In an experimental study of this relationship, the investigators could randomize the study participants by whether they did or did not use OCs. (This is one example of what we noted in Part I, that experiments in humans are often impractical and sometimes unethical.) If randomization is effective, we would expect the distribution of smokers and non-smokers to be approximately equal between OC users and non-users. The goal of randomization is to create groups of people who are equally likely to develop disease (myocardial infarction) in the absence of the exposure of interest (OC use). If this goal is achieved, potential covariates such as smoking will be distributed equally among the groups and will therefore have no effect on the exposure-disease relationship; the effect of smoking on the relationship will be controlled for.

Restriction ensures that potential confounders do not differ between study groups. For example, in a study of the relationship between OC use and myocardial infarction, any potential confounding effect of smoking could be controlled by eliminating smokers from the study population. However, if the study were restricted to non-smokers, the conclusions would not be generalizable because they may not apply to smokers.

The objective of matching is to ensure that covariates are equally distributed among study groups so that the groups are comparable with respect to the matching variable. For example, if the investigator matches on smoking, then smoking status is distributed the same among women who used OCs as among women who did not.

If successful, randomization, restriction, and matching will eliminate the impact of the covariate on the exposure-disease relationship. However, if the covariate is a potential risk factor of interest, it would be unwise to deal with it by one of these techniques. For example, if one wanted to determine whether the impact of OC use on the risk of myocardial infarction varied by smoking status (ie, whether smoking was an effect modifier), then restricting the study to non-smokers would preclude the ability to study the modifying effect of smoking. Therefore, study design strategies are used to eliminate the impact of covariates one is not interested in, so that one can evaluate an undistorted measure of the exposure-disease relationship in which one is interested.

Dealing With Covariates Through Study Analysis

If the impact of a covariate is not eliminated by study design, it must be addressed during study analysis. There are two approaches to the control of covariates during analysis: stratification and mathematical modeling.[1] In stratified analysis, study groups are categorized by relevant covariates. The association between the exposure and the disease is then evaluated for each category. For example, the risk of myocardial infarction for OC users could be calculated and reported separately for smokers and for non-smokers. Stratification can thereby provide a simple and useful way to identify the impact of covariates on the exposure-disease relationship. In fact, stratification is the best initial approach to controlling covariates during analysis. However, stratification is often impractical, particularly if more than one or two covariates must be controlled. Even large studies may have too few persons in each stratum to analyze. In such instances, when one "runs out of numbers," there is a resultant lack of both precision and reliability. For example, even if a study of OC use and myocardial infarction is large enough to stratify by smoking, it might be too small to stratify by age as well. In our example, the ability to stratify by multiple factors is important because the risk of cardiovascular disease associated with smoking and OC use appears to be modified by age. Multivariable mathematical models are needed if several covariates are to be controlled.

When selecting a mathematical model, the objective is to choose one whose properties and assumptions fit the study data as closely as possible. Because of its two main properties, the logistic model is often chosen to describe epidemiologic data. First, the logistic model provides estimates of risk for which values are restricted to a range between 0–1. "Risk" means probability, which always ranges between 0 (eg, no cardiovascular disease) and 1 (eg, cardiovascular disease). Therefore, when the logistic model is used, the resulting risk estimates describe probabilities. For other models, risk estimates above 1 or below 0 can occur. Second, the mathematical form of the logistic model is S-shaped. The biologic relationship between risk factors and the development of disease is often well described by an S-shaped curve.

We can illustrate the logistic model by using data from a completed analysis of the relationship between OC use and ovarian cancer,[2] in which both age and parity were covariates (in this case, they were confounders). Oral contraceptive use and the confounders, age and parity, are the independent variables we want to use to predict the disease, ovarian cancer. The investigator enters data obtained on the independent variables and on the disease into a computer by using an appropriate computer program for logistic regression. The program will then estimate a logistic model based on the data and provide relevant results on the computer printout.

For our example on OC use and ovarian cancer, the computer printout includes the following information:

Variable	Coefficient
Intercept	−1.1818
OC use (β)	−0.5336
Age	−0.0843
Parity	−0.7440

This type of information can be used to estimate an odds ratio, test for its statistical significance, and obtain confidence intervals around it. The latter concepts were introduced in Part I. Their application to logistic regression can be illustrated most simply if the logistic model contains the following: 1) a single 0/1 exposure variable (ie, one for which an individual is either not exposed [0] or exposed [1]), and 2) several potential confounders to the exposure-disease relationship, but no interaction terms (defined below). Given these two conditions, the formula for the estimated odds ratio (OR) is simply OR = e^{β}, where β is the estimated coefficient of the exposure variable calculated by the computer program.

In our example on the risk of ovarian cancer among OC users, the odds ratio adjusted for the confounders age and parity is estimated by e^{β} or $e^{-0.5336} = 0.6$. In other words, the logistic regression results tell us that, after we control for the effects of age and parity, OC users have a lower risk of developing ovarian cancer than women who do not use OCs. One might wonder why the coefficient β (the β estimate for OC use) is *adjusted for the effects of age and parity.* The answer is that the value of β depends on the coefficients for age and parity; the computer program calculates these interrelated coefficients. Standard logistic regression programs will also calculate the 95% confidence interval (see Part I) for the odds ratio.

To this point, we have illustrated only the use of logistic regression to control for confounders. Logistic regression can also handle the other major type of covariate, the effect modifier. Effect modifiers are dealt with by including product (interaction) terms in the logistic model. In our example, if age had been an effect modifier rather than a confounder (ie, if the impact of OC use on the risk of ovarian cancer varied by age), then an interaction term (eg, x = OC × age) would have to be included in the logistic model. The inclusion of interaction terms, which may even include variables raised to a higher power, can substantially complicate the building of a logistic model. Complex models are Discussed in introductory texts on regression analysis.

How Mathematical Models Are Built

To build a proper model for regression analysis, investigators must decide which covariates to include and in which sequence to enter or delete them. There are no uniform rules for this process. Although "cookbook" approaches to modeling carry some risks, a general strategy for model building has been proposed[1] that has proven useful. This strategy consists of the following steps:

1) Specify variables. Identify the exposure and disease of interest and the independent variables to be assessed as potential confounders. Identify interactions to be evaluated.

2) Construct an initial model. Include the exposure, the disease, potential confounders, and pertinent two-factor interactions (such as, in the example, OC use and age).

3) Assess interactions by statistical tests for product terms. If significant interaction is identified, calculate a different odds ratio for each category of the covariate in the product term. For example, if calculations show that the product term involving OC use (the exposure) and age (the covariate) is statistically significant, calculate odds ratio estimates for specific age groups of OC users. Thus, if there are four different age groups of interest, four different odds ratio estimates are calcu-

lated. In such a case, age, the effect modifier, need not be subsequently assessed as a potential confounder.

4) Assess confounding for those covariates not found to be effect modifiers. This is done by estimating the exposure-disease relationship with and without each potential confounder. If the estimates are meaningfully different, confounding is present. There are no rules for deciding what is "meaningful." Statistical testing is not used.

5) Draw conclusions about odds ratios of interest based on the final model. This model will contain appropriate confounders and interaction (product) terms together with the exposure variable(s) of interest.

Evaluating the Use of Mathematical Models

The critical reader may find it difficult to assess the appropriateness of regression analysis for a particular study. The methods section of a report often fails to explain why a particular choice of regression model is appropriate, whether confounders were identified a priori or by their impact on the exposure-disease relationship, or whether interaction was assessed. Direct communication with the authors of a report may be required to obtain this information. Such verification is clearly not feasible or practical on a routine basis.

So how should the practicing clinician evaluate a report in which logistic regression is used? Even when there is little information about the use of the model, there may be sufficient information to assess the quality of the data fed into the model (Part I of this report) and to draw meaningful interpretations from the odds ratios thus obtained. The use of logistic regression analysis suggests only that the data have been manipulated in a sophisticated, but not necessarily correct, manner. It must be emphasized that the use of mathematical modeling cannot "fix" data that result from a poorly designed or improperly conducted study. Nevertheless, when properly used, mathematical modeling can control for covariates and thereby improve the likelihood of obtaining an accurate measure of the exposure-disease relationship.

References

1. Greenberg RS, Kleinbaum DK. Mathematical modeling strategies for the analysis of epidemiologic research. Ann Rev Public Health 1985;6:223–45.
2. The Cancer and Steroid Hormone Study of the Centers for Disease Control and the National Institutes of Child Health and Human Development. The reduction in risk of ovarian cancer associated with oral contraceptive use. N Engl J Med 1987;316:650–5.

Address reprint requests to:
Herbert B. Peterson, MD
Women's Health and Fertility Branch
Centers for Disease Control, Mailstop K–34
1600 Clifton Road
Atlanta, GA 30333

Received November 26, 1990.
Received in revised form April 23, 1991.
Accepted April 23, 1991.